The Analysis of Journals Topics and Trend: Text Mining and Word Cloud

Jung Hoon Baeg
Florida State University

Abstract
The objective of this study is to identify the major topics over time of 4 selected journals in the field of Information Science and Library Science by analyzing word usage and the frequency with which certain words are used in the journal. A further objective is to determine the future direction of research in related subject areas. The basis of the study uses the program R to collect, text mine, and analyze a published article's word usage and concepts that are represented in a word cloud. Four journals were selected and collected from Web of Science (Thomson Reuters). The journal selection considered the 5-Year Impact Factor, the journal's aims and scope, and subject uniqueness. The journals that were selected are: *Journal of the Association for Information Science and Technology, Journal of the American Medical Informatics Association, Journal of Documentation*, and *Scientometrics*. A total 8,148 articles were collected and analyzed.
Keywords: Journal; text mining; R; word clouds; visualization
doi: 10.9776/16578
Copyright: Copyright is held by the author.
Contact: jhb6536@my.fsu.edu

1 Introduction

Every journal has its own aims and scope and each journal pursues them through peer-reviewed manuscripts. Each journal builds its own reputation and characteristics. The published articles may represent the journal's domain subjects. For example, the *Journal of the Association for Information Science and Technology* focuses "on the production, discovery, recording, storage, representation, retrieval, presentation, manipulation, dissemination, use, and evaluation of information and the tools and techniques associated with these processes (JASIST)." Manuscripts are peer-reviewed and selected not only based upon the aims and scope but also on the quality of the manuscript. However, a journal's aims and scope are one of the more important considerations when a researcher submits their work to the journal.

The topic and terms used in the articles may represent the journals subject domain. Funk's (2013) study showed how word usage in a body of literature identifies the article topic and trends over time. The more used terms indicate the topic of the articles, in the patent document (Noh, Jo, & Lee, 2015; Xie & Miyazaki (2013), and area of the Landscape and Urban Planning field (Gobster, 2014). Many studies use the text mining method to identify the most frequent term (key word) selection. Text mining allows researchers to retrieve information from different sources to merge and reorganize it for the purpose of study. Text mining is able to clean the data and eliminate duplication or unnecessary terms (Garechana, Belver, Cilleruelo, & Sarasola, 2014).

The purpose of this study is to determine how the terms (key words) used in each article reflect the topic of articles and the journal's scope and aims and also identify the characteristics of the journal. A further objective is to determine future research trends in Information Science and Library Science. The particular research questions are: How do all of the article's titles, abstracts, key words and usage of terms indicate *the topics* and each journal's subject domain?; Are there any word selections and usage that are different across the journals?; and Are there any indications of future research trends?

2 Methods

This study employs different methods and approaches. Data collections are from bibliometrics and term selection and data analysis are from text mining and information visualization.

2.1 Data

This study used data from the Web of Science (WoS) database. Web of Science is an online database maintained by Thomson Reuters that produce multiple database access and allow for cross-disciplinary research. They provide over 90 million records of which 5,300 are social science publications in 55 disciplines. They also provide the Journal Citation Reports every year.

To select the journals the 5-Year Impact Factor, a broad field of information science, and the aims and scope of the journal, and specific subject characteristics in the journal were considered. The four journals that were selected were the *Journal of Association for Information Science and Technology* (JASIST) (prior 2014, *Journal of the American Society for Information Science and Technology*), *Journal of the American Medical Informatics Association* (JAMIA), *Journal of Documentation* (JDOC), and *Scientometrics*. Journal articles were collected from from all of the journals between 2001 and 2015. The publication of each journal started in a different year, however to make the data consistent and search for recent research trends, the researcher select this period for data collection.

R is a free software for statistical analysis but also program language to use computing and graphical analysis. One of the strength of R is that it is free software and it is still developing so that it can be used for many research purposes. Many users develop 'packages' for specific analysis. In this study, R is used for text mining and graphical implementation to create word clouds. For the text mining the 'tm' package was used and the 'wordcloud' package was for word cloud. Word clouds allow the simple graphical implementation of each journal's unique research interest and trends but it also indicates approximate common interests

JASIST covered not only specific subjects (look for scope) but it also covers the broad areas of information science disciplines. JAMIA is a unique subject, medical informatics, which is one of the subject areas of LIS but JAMIA has amore specific research agenda. JDOC is a one of the longest running journals in library and information science A total of 8,148 articles were collected and the number of articles from each journal are as followed: JASIST (N=2848), JAMIA (N=1879), JDOC (N=1037), and Scientometrics (N=2680).

2.2 Text mining and word clouds

All data are from WoS saved as .txt file format. This data contains the author, title, source, and abstract. Sources consist of multiple abbreviations of the title and publisher, such as 'FN', 'VR', etc. It contains unnecessary terms to analyze. In order to accurately analyze the data, these unnecessary terms need to be removed. Program R package 'tm' is able to text mine for document.

In addition, R language is case sensitive, which makes this elimination possible. The important aspect of text mining, it [that it] needs to remove unnecessary words, numbers, punctuation using 'tm' packages removed. The command is 'tm_map(x, removeWords, removeWords), in addition using stopwords("english") use to remove common English words, such as 'for', 'very', 'of', 'are', and etc. The list of stopwords("english") consists of a total of 174 common words. An example of the instructions to remove English words is: doc <- (doc, removeWords, stopwords("english"). After text mining, the study made a term document matrix to list the frequency of words as the cells of the matrix. This matrix table will be used for data analysis.

Third, word clouds use the 'wordcloud" package, which is generated by the frequency of words based upon a term document matrix ('wordcloud(names(freq), freq, min.freq=100)). It effectively provides a quick visual overview of the results.

3 Results

3.1 *Journal of Association for Information Science and Technology* (JASIST)

The most frequently used terms are 'information', 'web', 'data', 'retrieval', and 'knowledge' in JASIST. These terms indicate the subject domain (topic) in JASIST and those domains also match with the theme of the journal. The word cloud showed the big letters with the majority of the topics used in this journal (Figure 1). We also interpreted JASIST to be technology implementation driven by 'web' related topics in the journal.

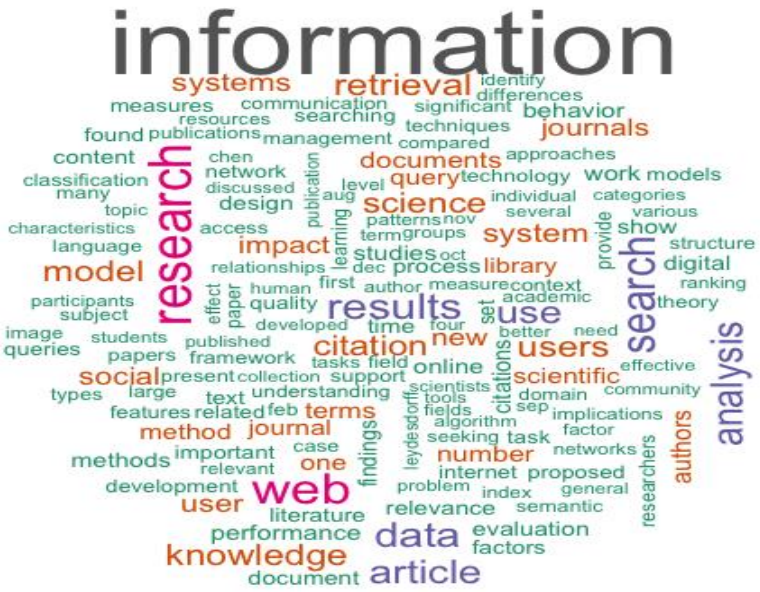

Figure 1. JASIST

3.2 *Journal of the American Medical Informatics Association* (JAMIA)

JAMIA focuses on biomedical and health informatics in the area of "clinical care, clinical research, translational science, implementation science, imaging, education, consumer health, public health and policy (JAMIA)." Figure 2 showed 'data', 'clinical', 'health', 'information', 'patient', significant. The majority of the topics are these terms related. In addition 'electronic' and 'systems' also determine implementation science that is one of focus on JAMIA (Figure 2). The majority of topics are suitable in the JAMIA.

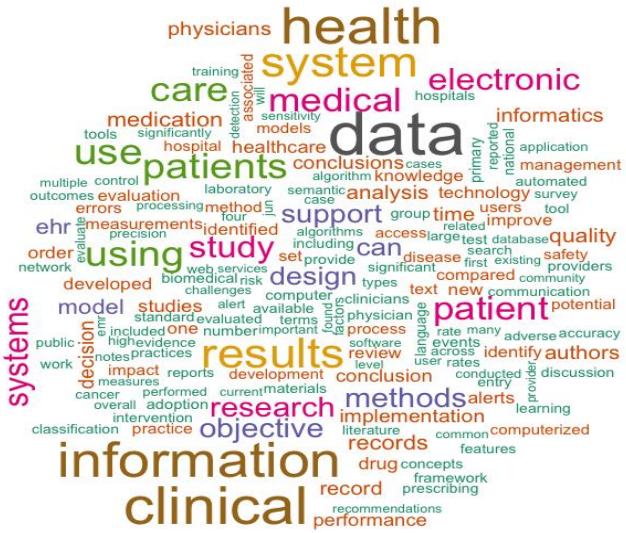

Figure 2. JAMIA

3.3 *Journal of Documentation* (JDOC)

JDOC is the one of the longest academic journals in Information Science and Library Science. JDOC focuses on "theories, concepts, models, frameworks and philosophies related to documents and recorded knowledge (JDOC)." The word cloud showed the majority of topics, such as 'information', 'documentation', 'design' and 'methodology' match with the journal's own focus, (Figure 3).

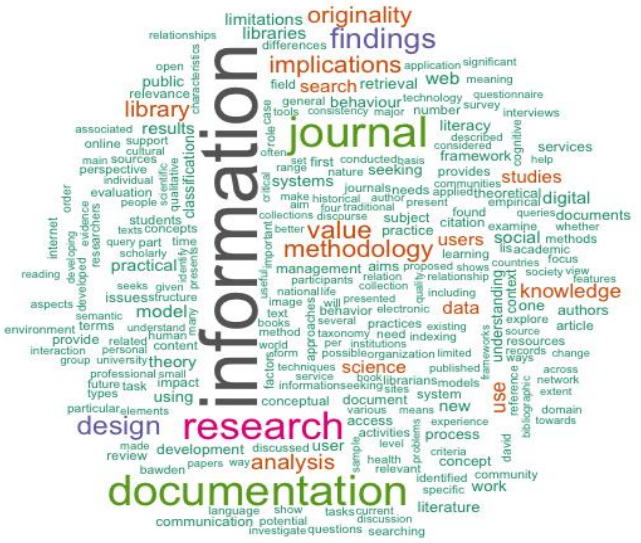

Figure 3. JDOC

3.4 *Scientometrics*

Scientometrics focuses on the “quantitative features and characteristics of science and mathematical (statistical) methods”, such as scientometrics and related fields. The articles main topics are driven by mathematical methods, such as ‘science’, ‘citation’, ‘data’, collaboration’, and ‘analysis’ (Figure 4). These topics are broadly part of the domain in bibliometrics. Scientometrics journals are well known for bibliometrics as well.

Figure 4. Scientometrics

3.5 Four journals comparison

The comparison of the four journal word clouds showed the common topics as well as each journal’s main subject interest. ‘Information’ is the biggest characteristic and it indicates the most common topics among all of the journals. ‘Documentation’ and ‘journal’ are also remarkable as they indicate two topics, which are common and show a common interest between Scientometrics and JDOC. The comparison word cloud showed that each journal has its own domain subject clearly articulated (Figure 5).

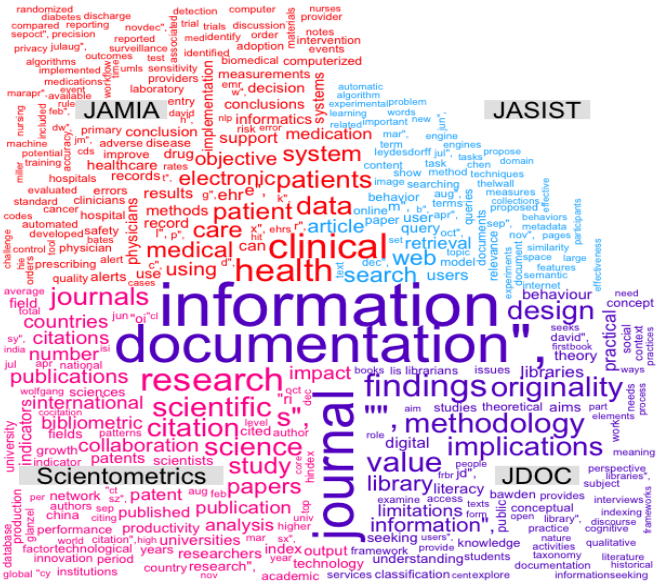

Figure 5. Journals comparison

4 Conclusion

Text mining and word clouds showed each journal's major topic of interest. The results may be obvious because each journal pursues their own aims and scope through peer-reviewed manuscripts. However, this study empirically proved that all of the four selected journals have their own interests as well as shows that there are topic differences. Another important finding was all four journals have common a topic, which is 'information' related subjects indicating that all four journals are deal with subjects in the discipline of Information Science and Library Science. This study has some limitations. Mainly, the data was collected over a certain period time, which might mislead the majority topics over time. A future study needs to be done with all published articles.

5 References

Funk, M. E. (2013). Our words, our story: a textual analysis of articles published in the Bulletin of the Medical Library Association/Journal of the Medical Association from 1961 to 2010. *Journal of the Medical Library Association, 101*(1), 12-20.

Gobster, P. H. (2014). (Text) Mining the LANDscape: Themes and trends over 40 years of landscape and urban planning. *Landscape and Urban Planning*, *126*, 21-30.

JASIST [n.d.]. Retrieved June 2, 2015 from JASIST homepage:http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%2915322890/homepage/ProductInformation.html

JAMIA [n.d.]. Retrieved June 2, 2015 from JAMIA homepage:http://jamia.oxfordjournals.org/for_authors/index.html

JDOC [n.d.]. Retrieved June2, 2015 from JDOC homepage: http://www.emeraldgrouppublishing.com/products/journals/journals.htm?id=JD

Scientometrics [n.d.]. Retrieved June 2, 2015 from Scientometrics homepage:http://www.springer.com/computer/database+management+%26+information+retrieval/journal/11192

Van Eck, N. J., Waltman, L., Noyons, E. C. M., & Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics, 82*(3), 581-596.

Xie, Z., & Miyazaki, K. (2013). Evaluating the effectiveness of keyword search strategy for patent identification. *World Patent Information, 35*(1), 20-30.

Yuno Do, J. S. (2014). Research topics and trends over the past decade (2001-2013) of Baltic Coleopterology using text mining methods. *Baltic Journal of Coleopterology, 14*(1), 1-6.